Karlsruhe Institute of Technology

Japanese-English Machine Translation for a Humanoid Robot Moderator

Student Research Thesis of

Henning Sperr

At the Department of Informatics
Interactive Systems Labs
Institute for Anthropomatics

KIT:	Prof. A. Waibel
Waseda:	Prof. T. Kobayashi
Advisor:	Jan Niehues
Advisor Waseda:	Shinya Fujie

Duration: 01. May 2012 – 31. July 2012

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

PLACE, DATE

.....

(**YOUR NAME**)

Contents

1. Zusammenfassung

Diese Arbeit entstand zum Teil an der Waseda Universität Japan und besteht aus zwei Teilen. Wir haben versucht ein Roboter System zu entwickeln das im Vorlesungskontext erkennen kann ob Übersetzungsprobleme bestehen und eine Übersetzung anbieten kann. Die Ausgangsidee war, dass Japaner häufig Englisch besser verstehen können als selbst zu sprechen. Darum wäre ein Roboter hilfreich, der in einer Vorlesung als Unterstützer zur Seite steht und versucht die Kommunikation zwischen Studenten und Vortragenden zu verbessern. Der Vortragende hält in diesem Fall eine Vorlesung auf Englisch und die Studenten sind Japaner. Wenn einer dieser Studenten eine Frage hat und versucht mit dem Vortragenden zu kommunizieren, versucht der Roboter zu erkennen ob Missverständnisse entstehen und gegebenenfalls einzuschreiten und seine Übersetzerfähigkeiten anzubieten. Dies soll sich so anfühlen wie der "gute Freund" der in der Vorlesung neben einem sitzt und hilft dem Lehrer zu erklären worum es geht. (Abbildung 3.1) Der Hauptteil befasst sich mit Phrasenbasierter Statistischer Maschineller Übersetzung, einer der vielversprechendsten Ansätze zur Maschinellen Übersetzung. Es wurde untersucht wie gut sich mit vorhandenen Mitteln ein geeignetes System für Japanisch-Englisch herstellen lässt. Um gesprochene Sprache zu übersetzen benutzen wir dabei die öffentlich verfügbaren TED Untertitel und versuchten auch mit zwei anderen Korpora die Übersetzungsgenauigkeit zu verbessern. Zunächst haben wir versucht mit verschiedenen Tokenizern, Programme die Wortgrenzen erkennen und kenntlich machen, die jeweils eine andere Segmentierung der Sprache vornehmen den besten BLEU score zu erzielen. Anschliessend haben wir versucht verschiedene Umordungs-Techniken zu verwenden um die Unterschiedliche Satzstellung zwischen Englisch und Japanisch auszugleichen. Desweiteren haben wir untersucht mit verschieden Heuristiken der Phrasenextraktion bessere BLEU Scores zu erzielen. Im zweiten Teil der Arbeit ging es darum den Roboter zu entwickeln und den Spracherkenner und Übersetzer zu integrieren. Weil das Erkennen des Inhalts von Konversation ein schweres Problem ist, an dem gegenwärtig noch Forschung getrieben wird, haben wir uns darauf beschränkt ein einfaches System was gegebenenfalls Probleme erkennen kann zu konstruieren. Das entwickelte Programm benutzt den Spracherkenner und versucht festzustellen ob der gegebene Input eine Hesitation (z.B.: Ähm, Öhm) ist. Aus der Anzahl der Hesitationen zwischen zwei gesprochenen normalen Sätzen wird versucht zu schliessen, ob der Sprecher gerade Probleme hat sich auszudrücken. Um Hesitationen zu erkennen haben wir unterschiedliche Metriken implementiert. Je nachdem was der Nutzer antwortet, wird entweder der Übersetzungsprozess gestartet oder von neuem gehört ob es Probleme geben könnte. Die Übersetzung wurde so konzipiert, dass der Übersetzer als ein Webserver in Karlsruhe läuft und von Waseda jederzeit eine Übersetzung erfragt werden kann.

2. Abstract

This work was created partly at Waseda University Tokyo, Japan and consists of two parts. We have tried to develop a robot communication system which can recognize communication problems in lectures. If such problem is detected the system should be interacting with the students and lecturer and offer translation help. The idea behind it was that Japanese people often times understand the English language better then they can speak it. That is why a robot is helpful that could act as a supporting friend in English lectures to Japanese students who try to ask questions or start a discussion. If a student tries to start discussion the robot should monitor the conversation as third party and try to interfere only if it thinks there might be some benefit in creating clarification or aid translation difficulties. This should create the feeling of the "friend" sitting next to you in a lecture. (See Figure 3.1) The main part consists of phrase based statistical machine translation, which is one of the most promising approaches in machine translation. We used known technologies to see how good they work with Japanese English translation. To be able to translate spoken language we tried to use the TED open video lecture subtitles to get good data for parallel spoken language corpora. We also used two other freely available corpora to increase the translation accuracy. First we concluded tests on the tokenization of Japanese language. Since Japanese language does not contain any spaces by itself this is a nontrivial task and there are different approaches in doing so. We also tried different ways of extracting phrases to increase BLEU scores. Several reordering techniques which should balance the different sentence orderings in Japanese and English language. In the second part of my work, we tried to develop the robot so it can recognize speech and trigger the translation process if needed. Since recognizing the topic of conversation is a very hard task for computers and subject to recent research we decided to build a very simple method of recognizing conversation flaws. Our system tries to listen for hesitations of the speaker (e.g. Well... Umm...) and tries to see if the participant is unsure on how to express himself. Since in our task it will be Japanese students trying to speak English, we can conclude that long thinking with a lot of hesitations might be an indicator for translation problems. The robot then asks the student if he needs translation aid and provides it if necessary.

3. Introduction

The task at Waseda University was to create a robot which aids communication in lecture context. A real life scenario that can be found at a lot of Japanese universities is that English speaking teachers hold a lecture in front of Japanese people. Japanese people are often able to understand the English language but it is hard for them to speak it themselves. So there might be students who have a question about the topic of the lecture but do not dare to ask questions. Since they do not know how to phrase the question in English. Maybe they are self-confident and try to express it in English but they hesitate or create very broken sentences so it is hard for the lecturer to understand what the question of the student is. The lecturer might have the same problem speaking Japanese and misunderstandings or no communication can happen. That is why we implemented a robot, that will listen to the conversation of student and lecturer and try to find out if the quality of the communication is good. If the robot detects problems or thinks that translation might be needed he will ask the person and offer its translation skills. After the translation the robot goes back to listening mode and begins to observe again. (Figure 3.1)

The difference to a simple computer with an interface is that the robot will give people the opportunity to try by themselves and will participate in the conversation in a natural way. Like the classmate who is sitting next to you who is helping you out. Since this is a very broad field with many different problems we started working on it with the focus on the translation approach. For the detection of misunderstandings we just created a simple demo scenario showing the principle of the system. And how it is supposed to work.

3.1 Statistical Machine Translation

The beginning of machine translation reaches back to the cold war, where the Americans tried to spy on Russian radio transmissions. The first approaches saw translation as a form of decoding a encrypted document, which contains English encrypted into some other language. In 1966 the so called ALPAC report stalled the progress of research for nearly 20 years. ALPAC[1] concluded that machine translation was an unfeasible task. In the 1980s big companies like Siemens (Metal Project) began to research again and laid the first stones for statistical machine translation. In the early 1970s the first rule-based translation systems were developed. Rule-based Systems use dictionaries and grammatical

[1] http://www.nap.edu/openbook.php?record_id=9547

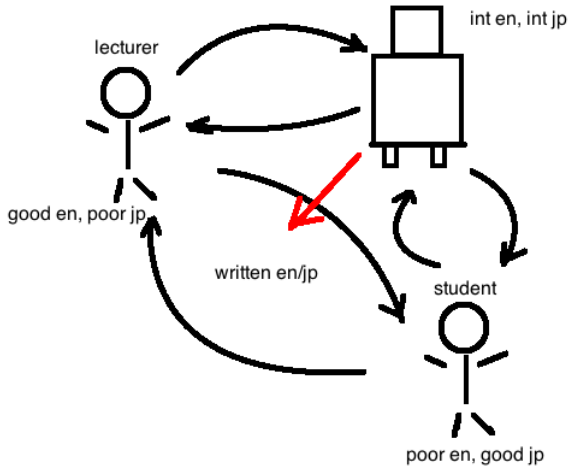

Figure 3.1: Conversation Robot

rules to translate one language into another. There are still systems from this time, like the now hybrid rule-based/statistical system called "Systran", which is still commercially developed. In the early 1990s IBM developed a statistical word based translation approach. They used big parallel corpora to learn word translations in different contexts. The new and good thing was that they did not need to have any linguistic knowledge of the target language. This system has been developed in several so called IBM Models. There has been a lot of progress since the early approaches of Brown et al. (1993), who worked for IBM research. Newer systems use the phrase based translation approach (Koehn at al 2003 [KOM03]). Compared to the IBM Models the smallest translatable unit is a phrase, which can consist of one or several words. This helps to improve translation quality since it is possible to translate multiple words at once. The first few IBM Models are still used in the phrase-based approach to find an alignment between two corpora. This means the algorithms for the IBM Models can generate a probability alignment between the words of two parallel corpora. With this alignment there are different heuristics how to extract phrases. After you extracted the phrases you simply count how often you saw each phrase and assign its probability. Statistical machine translation is one of the most promising approaches to conquer automatic translations. Recent development is looking how to improve the phrase based approach by using part of speech and different reordering algorithms.(Lioma et al. 2005[LO05]) In this thesis we are building a phrased based Japanese-English system.

3.2 Tools

There are some freely available tools from different researchers on the internet with which it is possible to build a statistical translation system. For our work we used following tools:

GIZA++ (Och 2000, [ON00]) is an implementation of the IBM Models. Today it is used to find the most probable alignment between two corpora.

Moses[2](Koehn et al. [KHB+07]) is an implementation of a phrased based decoder from the University of Edinburgh. We did not use moses for decoding but a few of its scripts to aid our preprocessing.

[2]Available at: http://www.statmt.org/moses/

SRILM (Stockle [Sto02]) is a language modeling toolkit which we use to build our language model.

STTK (Statistical Translation Toolkit[VZH+03]) Is the decoder we used. It is developed at the Carnegie Mellon University.

BLEU (Bilingual Evaluation Understudy) is an algorithm for judging correctness of machine translated text. BLEU compares the translated text to a human translated reference and evaluates how close the machine translation is. BLEU was designed to approximate human judgement at a corpus level.

MeCab[3] is a Japanese morphological analyser designed for generic purpose natural language processing tasks. It can convert Kana to Kanji, generate part-of-speech tags and separation of Japanese word.

Kytea[4] is another Japanese morphological analyser. It uses different methods then MeCab to generate word segmentation and part-of-speech tags(Neubig et al.[NNM11]).

IPADIC is a Japanese word dictionary data designed to be used for the morphological analysis system to segment and tokenize Japanese text string into unit words. It provides many additional information (pronunciation, semantic information, and others).

3.3 Corpora

Statistical machine translation systems require a lot of data to make good predictions. It is hard to find free parallel corpora that are also of good quality. A lot of institutes have paid people to do a lot of translation work but this is expensive and very time consuming. We found three sources for parallel Japanese and English texts. The quality of the data is not perfect since sometimes it is very loose translations or it is about a really specific domain. We did experiments as we discovered the corpora so some of the later experiments are for example done only on TED and not on KFTT anymore. Only the most promising of approaches have been conducted on all corpora since the lack of time.

3.3.1 TED Data

TED is a huge nonprofit internet platform with talks about "ideas worth spreading"[5]. There are many talks online and in the recent times TED organization started to allow subtitling of their talks. They provided the English subtitles for the videos and users are able to translate them into any other language. Right now there are around 700 talks for which English and Japanese subtitles exist.

We tried to use this data to learn how to translate spoken Japanese language inside the research domain. Since TED only released the English subtitles and the translation was done by the viewers some of the translations are translating the meaning instead of the precise text, which makes it hard for a general translation system, to translate the right reference. Also in Japanese language leaving information out of the sentence because it should be clear from the context creates sentences which are hard to translate.

It right now contains 74.768 parallel sentences with around 1000 sentences for development and test set

[3]Available at: http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[4]Available at: http://www.phontron.com/kytea/
[5]http://www.ted.com/pages/about

3.3.2 Kyoto Free Translation Task

The Kyoto Free Translation Task[6] (KFTT, Neubig 2011 [Neu11]) is a task for evaluation and development of Japanese-English translation systems. It is a freely available corpus of Wikipedia articles related to Kyoto. Its focus is to measure the performance of Japanese-English translation systems. The data was originally prepared by the National Institute for Information and Communication Technology (NICT) which released it to public domain under the common share alike licence. The direction of Japanese to English was checked and translated by professional translators while the other direction is difficult an still needs a lot of work. The type of text is encyclopedic and it is in a specialized domain, namely Kyoto.

It contains 440.288 parallel sentences with 1200 sentences for dev and test set.

3.3.3 Tatoeba Corpus

The Tatoeba[7] corpus started from the Tanaka corpus which was released into public domain in 2001 from Professor Yasuhito Tanaka at Hyogo University. He and his students compiled a list of 300 parallel sentences per student in English and Japanese every year. After several years the corpus had grown to 212.000 sentences. The domain of this corpus is sentences from various English study books, songs, other popular books or even the bible. In the begging the corpus was full of errors in the translation as in the Japanese itself, sometimes the translation did not even match at all. After the corpus got included in the WWWJDIC[8] it got cleaned and duplicates got removed which reduced the size to 180.000 pairs. Sentences which only differed by orthography, for example using Kana, Kanji, numbers or proper names were combined to one representative example. After all the cleanings the corpus remained with roughly 140.000 sentences. In 2006 the Tanaka Corpus was incorporated into the Tatoeba project, which is a huge database of sentences translated into several languages. It is a platform to help language learners to learn a language by learning translations of sentences. You can download and obtain the parallel sentences from Tatoeba by filtering the language pair you want. And extracting it from the Database file.

The corpus contains 137.716 sentences and 2000 for dev and test set each.

3.4 Japanese Language

Japanese language is made up of three different writing systems. They have the Kana which is divided in Hiragana (48 characters) and Katakana (48 characters) and they have the Kanji (around 2000+) which are the chinese symbols. Each Hiragana and Katakana character denotes one mora, which is a phonological unit for example ”ka,ke,ki,ko,ku” almost all Japanese mora consist of either a single vowel or a consonant followed by a vowel. For each Hiragana there is also a Katakana character. Katakana was introduced to write words that came from abroad like インタネト (intaneto - Internet). Kanji are small pictographs originating in China. The reading of a Kanji can not be guessed by its pictograph. Each Kanji has so called Onyomi and Kunyomi which denotes two different readings. Onyomi translates to ”Sound reading” which denotes the original reading from china. Japanese people tried to approximate the chinese reading with their own language. Often there are several different Onyomi since the Kanji was introduced to Japan several times in different regions and different years and Japanese people kept most of the readings. Kunyomi which means ”meaning reading” is the Japanese pronunciation of the Kanji.

[6] http://www.phontron.com/kftt/
[7] http://tatoeba.org/eng/
[8] http://www.csse.monash.edu.au/jwb/wwwjdicinf.html/# example_ tag

Japanese people associated the chinese Kanji with the word for their pictographs meaning and adapted that reading aswell.

Japanese language has no spaces between words which creates difficulties for the machine translation process. For the alignment of words between parallel corpora it is necessary to have spaces between word boundaries. This process called tokenization is a problem in translation between Japanese and English. There are several methods on tokenizing as we will see in Section 4.1.

Another problem with Japanese to English translation is that Japanese language has a subject object verb (SOV) word order while English as most european languages has a subject verb object (SVO) word order. This can create the necessity for long word reorderings which normally decoders are unable to do themselves, hence requiring a better treatment of the reordering.

Spoken Japanese is highly context sensitive, which means that the speaker does not say anything that can be guessed by the context. A correct Japanese sentence can sometimes just be a single verb. So depending on the context and intonation the sentence with just the verb 食べたい (tabetai - want to eat) can have the meaning ”I want to eat.”, ”he wants to eat”, or ”do you want to eat?”. This is really hard for machine translation systems, because they do not know the context the speakers are in. So starting to build a translation system is easier on written Japanese, where the sentences are well formed and nothing is left out.

4. Translation System

For our Baseline System we used the Systembuilder of the Institute of Cognitive Sciences at the Karlsruhe Institute of Technology. The Systembuilder is a very flexible python script combining all the steps from preprocessing (e.g. see Section 4.1), to running GIZA++ for the IBM alignment and finally being able to run the STTK Decoder.

Since the main goal is the lecture translation system, which is in the spoken language domain, we decided to use the the TED open lecture subtitles. We tried different methods of tokenizing the Japanese language. Furthermore we were continuing our search for freely available corpora to aid our translation system and get better statistics for Japanese sentences. In the end we obtained and preprocessed the Kyoto Free Translation Task which was later dropped again since the Japanese is very old and in the wrong domain. We also found the Tatoeba database which contains lot of parallel sentences that should help people to learn a foreign language. Tatoeba should later become the main corpus of the translator since the TED data is prone to errors and contains free translations, instead of literal translations.

Using Tatoeba corpus we researched how combining and adaption of phrase tables would effect BLEU scores. For adaption we would train a translation model on all the available data and then train a separate in-domain model on the in-domain data (e.g. Tatoeba) reusing the same alignment from the large model. The two models are combined afterwards with a log-linear combination to get adaption towards the target domain. The finished model is then using four scores, the scores from the general model and the two smoothed relative frequencies of both directions from the i- domain model. If a phrase pair is not in our in-domain model, we use a default score instead of the relative frequency. We could not find an significant increase of scores no matter on which development and test set we ran the scoring and in which way we adapted the phrase tables.

Next bigger task was to research if learning reordering rules would improve BLEU results. We tried two different reordering approaches. One is to reverse the tokens of the Japanese sentences according to special rules and the other was to learn general reordering rules from the alignment of the IBM models. First method did not show to increase BLEU where as second method turned out to increase the BLEU score by around one point.

4.1 Tokenisation

Japanese language has no spaces between words, but for Statistical Machine Translation it is important to separate the words to find an alignment and extract phrases. According to Graham Neubig the Author of the Kyoto Free Translation Task and Kytea there are two major ways to tokenize Japanese language. [1]

1. Dictionary-based sequence-prediction methods
2. Word-boundary prediction methods

4.1.1 Dictionary-based sequence-prediction methods

For dictionary-based sequence prediction methods a huge dictionary of words combined with part of speech (POS) information is created. Tools like MeCab or its predecessor ChaSen use this approach. The Japanese text is parsed to find the best sequence of words in the dictionary. To determine the best score either hand assignment (e.g. JUMAN[2]) is used or it is evaluated by statistical methods like Hidden Markov Models or Conditional Random Fields (Lafferty et al [LMP01]).

4.1.2 Word-boundary prediction methods

In word-boundary prediction methods no dictionary for scores is needed. For this reason word-boundary models are easier to evaluate. The idea is to predict how likely it is that between two adjacent words a boundary exists in-between. Examples for this approach are TinySegmenter and Kytea(Neubig et al.[NNM11]).

4.1.3 Experiments

Our baseline system is the following: First we tokenized the texts, and generated alignments using GIZA++. With the alignment we would create a phrase table using the grow-diag-final-and heuristic. Together with a four gram language model we use the STTK Decoder.

To find the best tokenizer for Japanese language we used above explained baseline system and used different tokenizers. The results are shown in Table 4.1 where we used three different Japanese language tokenizers. For Kytea we used different probability models that are obtainable at the Kytea homepage[3]. Modelsizes ranged from 13Mbyte, 30Mbyte and 140Mbyte and are stated with different word accuracy. Since the results of the different models did not affect the BLEU score significantly, the table only shows the highest score of all Models, achieved with the 30Mbyte version which will be used in the following.

Configuration	BLEU
BaselineKytea	4.43
BaselineMeCab	4.94
BaselineTinySegmenter	4.23

Table 4.1: Baseline on TED data using different segmentation.

[1] http://www.quora.com/What-are-some-Japanese-tokenizers-or-tokenization-strategies
[2] Available at: http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN
[3] http://www.phontron.com/kytea/model.html

In Table 4.2 is an example how Kytea and MeCab tokenize a sentence, and how a human would do it. This shows the different possibilities of tokenizing a Japanese sentence. Even amongst different humans you can find different tokenizations.

Kytea	今日 は エネルギー と 気候 に つ い て 話 そ う と 思 い ま す 。
MeCab	今日 は エネルギー と 気候 について 話そ う と 思い ます 。
TinySeg.	今日 は エネルギー と 気候 について 話 そう と 思い ます 。
Human	今日 は エネルギー と 気候 について 話そう と 思 います 。

Table 4.2: Sentence of TED data in different tokenization.

4.2 Phrase Extraction

Normally phrases are extracted creating a bidirectional IBM alignment with GIZA++ and combining these alignments using heuristics. With phrase-based translation usually both alignments ”target-to-source” and ”source-to-target” are generated and then combined. The easiest possible way is to either use one of those alignments without combining and extract phrases from that. Another heuristic is to create the ”union” or ”intersection” of both alignments and extract phrases from that. A more sophisticated approach is ”grow-diag-final-and”. It starts with the intersection of both alignments and searches if neighboring words are also aligned in the union of both alignments and adds these alignment points. After that it takes all remaining points in the union that are not yet aligned and aligns them if neither the target nor the source word is aligned.

For most of the western languages the strategy ”grow-diag-final-and” showed to be the best strategy for creating phrase extraction alignment. It was not clear if this also holds for Japanese language, so we researched if it is possible to increase the translation score by using a different combination strategy. We tried using ”grow-diag-final-and”, ”intersection”, ”union”, ”source-to-target” and ”target-to-source”. We did this test before we discovered the MeCab tokenizer so it was only done on Kytea preprocessing.

In Table 4.3 we see the results of the baseline system using Kytea tokenizing to test different heuristics for extracting the phrases.

Configuration	BLEU
TED-grow diag final and	4.33
TED-union	3.42
TED-intersection	**4.55**
TED-target to source	*4.37*
TED-source to target	3.93
Kyoto-grow diag final and	**15.90**
Kyoto-union	11.91
Kyoto-intersection	13.70
Kyoto-target to source	12.49
Kyoto-source to target	12.03

Table 4.3: Phrasetable tests using Kytea segmentation

We concluded that ”grow-diag-final-and” also holds the best performance since intersection method is only slightly better on TED but it takes over 20 hours to extract the phrases. This is due the fact that intersection has the most unaligned points and leaves the most space for taking different phrase boundaries.

4.3 Reordering the Sentences

The decoder has a distortion model which usually gives a bad score to sentences with too long reorderings. Since Japanese in general has a ”subject object verb” (SOV) word order while English language has ”subject verb object” (SVO) it is necessary to find a method to do long range reorderings. We tried two different approaches in preprocessing.

4.3.1 Syntactic Reodering using hand-made rules

The first method tested was by Katz-Brown et al [KBC08].

In this approach the Japanese sentence gets split at the topic marker particle は . Then the sentence is split into two parts, the one before and the one after the topic marker particle. In the next step each of those parts is reversed and then the sentence will be combined again. In Table 4.4 we can see an example for a reordered sentence. The part before the topic marker は does not change since it is only one word. The part after the topic marker gets reversed and appended again. The position of 。 does not change in this procedure.

Tokenized	今日 は エネルギー と 気候 について 話そ う と 思い ます 。
English	Today engery and climate about talk going think.
Reordered	今日 は ます 思い と う そ 話 について 気候 と エネルギー 。
English	today think going talk about energy and climate.
Reference	Today I’m going to talk about energy and climate.

Table 4.4: Reordered Result.

Since in Japanese and English the topic of the sentence is at the beginning we want to leave the topic at the beginning of the sentence. The only thing that will be changed is that the Japanese verb, which is at the end of the sentence will now be in the middle and the objects at the end of the sentence. The proposed Method has two main reorderings. First it moves the verbs from the end of the sentence into the middle 話そ う と 思 い ま す (translation: talk think) bringing it closer to English word order.

Second it does local reoderings す ま い 思 と う 話そ which helps translation since in English the auxiliary words often precede the verb they assist, while in Japanese auxiliaries and inflections follow their verb.

In our tests the above proposed method did not show to increase the BLEU score on TED or Kyoto, we did not conclude experiments on Tatoeba.

4.3.2 Reordering using learned Part of Speech-based rules

The Method used was introduced by Rottman and Vogel [RV07] and Niehues et al.[NK+09]. They use part of speech (POS) information to extract Long-Range and Short-Range rules. The Short-Range rules are only reordering continuous segments while the Long-Range rules reorder discontinued parts of the sentence. (Example below) The algorithm uses the alignment function and the POS information to find general reorderings of certain word types. With this information a lattice (See Figure 4.1) is created which only contains the most probable of reorderings. This lattice is used to reorder the sentence before decoding and trying to find the best translation using different paths though the lattice.

In this method we learn rules that look like Table 4.5. In the first part we see the part of speech pattern the rule is looking for. It searches a sentence starting with a noun (N) and then a predicate (P) then there are some arbitrary number of words and then predicate

again then a verb (V) and an auxiliary verb (AV). Whenever we find this pattern we generate a rule that reorders this pattern given in the next row. The last two words of this pattern will be reordered to the front the rest will follow in order and all the words covered by —X— will wander to the back of the sentence. The resulting order is seen in the second line. The last column gives us the probability for that reordering to happen.

POS-Tags		Reordering	Probability
Rule:	N P ———— X——— P V AV	2 3 4 0 1	0.00337729
Reordered:	V AV N P P ———— X———		

Table 4.5: Reordering Rule Example.

In Table 4.7 you can see the different reordering rules learned for our example sentence. Rules one and three have multiple applications on the sentence. In the second column of the Table you can see the sentence and its Part-of-Speech tags. Beneath is the matching of the rules to the POS tags. In the fourth column you see how the rules reorder the given tags and in the last column you see the probability for this rule to be applied.

In Table 4.6 we can see how the rules applied by the decoder yield in different translations depending on which rules we allow. Long-Range rules include all rules found Long- and Short-Range rules, whereas Short-Range rules only contain rules that do reorderings within continuous segments. What can be seen is that in our example using no or only Short-Range reorderings results in worse sentence structure than also including the Long-Range rules. What Long-Range reordering still fails in this case is to bring ”I think, I will talk about” to the middle of the sentence.

Sen	今日 は エネルギー と 気候 に つ い て 話 そ う と 思 い ま す 。
REF	I am going to talk today about energy and climate .
HYP Long-Range	today , is the energy and climate , I will talk about , I think .
HYP Short-Range	today , it is energy , climate change , and I will talk about , I think .
HYP No-Reordering	today , energy and I will talk about climate , I think .

Table 4.6: Reordering in preprocessing example.

lit transl.	Today		energy	and	climate	about	talk			think	to do	.		
JP	今日	は	エネルギー	と	気候	について	話そ	う	と	思い	ます	。		
TAGS-JP	名詞	助詞	名詞	助詞	名詞	助詞	動詞	助動詞	助詞	動詞	助動詞	記号		
TAGS-EN	N	P	N	P	N	P	V	AV	P	V	AV	SB		
	Rules													
1	N	P	——	–X–	——	P	V	AV					2 3 4 0 1	0.00337729
1.1	N	P	——	–X–	——	——	——	——	P	V	AV		2 3 4 0 1	0.00337729
2	N	P	——	–X–	——	P	V	AV	P	V	AV		2 3 4 5 6 7 0 1	0.00395257
3	N	P	——	–X–	——		V	AV					2 3 0 1	0.00184518
3.1	N	P	——	–X–	——	——	——	——	——	V	AV		2 3 0 1	0.00184518
4	N	P	——	–X–	——	——	V	AV	P				2 3 4 0 1	0.0013371
5	N	P	——	–X–	——	——	V	AV	P	V			2 3 4 5 0 1	0.00196935
6	N	P	——	–X–	——	——	V	AV	P	V	AV		2 3 4 5 6 0 1	0.00192422

Table 4.7: Reodering Rules learned.

16

4.4 Results

Table 4.8 shows the final BLEU results achieved. We used the baseline system with MeCab tokenzation including the reordering using POS tags. The BLEU score on TED is very low which is to be researched why. In Table 4.9 and Table 4.10 are some example sentences for the quality of translation. While on Tatoeba the general translation accuracy and quality is much better then on TED there are still problems with the reordering. Further experiments can be conduced researching the quality of the reordering rules. Another way of improving the reordering might be to use bigger lattices, this means including more improbable reorderings aswell. On TED the overall quality of translation is very bad which might be because it is translation of spoken speech and often times the translation provided by the TED community is often not literal translation. Another possibility might be that it is not enough data for training. On TED reordering didn't have a positive effect. Research needs to be done on how good phrase alignments are and if for example discriminative word alignment can yield a increase in translation quality.

Configuration	BLEU
TED Baseline	**4.68**
TED Long Range	4.47
Tatoeba Baseline	15.21
Tatoeba Long Range	**16.44**
Tatoeba Long Left	15.44
Tatoeba Long Right	16.28
Tatoeba Short Range	14.91
Both Baseline	14.96
Both Adapt Tatoeba	14.96
Both Adapt Tatoeba Long Range	**15.81**

Table 4.8: Final Results on Tatoeba and TED

Sentence	今日 は エネルギー と 気候 について 話そ う と 思い ます 。
Reference	I am going to talk today about energy and climate .
Hypothesis	today , is the energy and climate I will talk about , I think .
Sentence	そこで もう 一つ 皆さん が 驚く だろ う もの を お 見せ し たい と 思い ます 。
Reference	and so I want to show you one other thing that may catch you a little bit by surprise .
Hypothesis	so another one of you , it would be something that I want to show you , I think .
Sentence	どう やっ て と 僕 は 尋ね た 魚 の 専門 家 に なっ た ん だい ？
Reference	I asked him how he became such an expert on fish .
Hypothesis	how do you , and I asked the fish experts in ?

Table 4.9: Example sentences on TED

Sentence	始発 列車 に 遅れ ない よう に 、 彼女 は 朝 早く 家 を 出 た 。
Reference	she left home early in the morning for fear that she might miss the first train .
Hypothesis	so as not to be late for the first train , she left home early in the morning .
Sentence	その ニュース を 聞い た とき 、 とても 悲しく 感じ た 。
Reference	I felt very sad when I heard the news .
Hypothesis	I heard the news when I felt very sad .
Sentence	私 は 彼 の 欠席 の 理由 を 知り たい 。
Reference	I want to know the reason for his absence .
Hypothesis	I want to know him in the absence of reason .

Table 4.10: Example sentences on Tatoeba

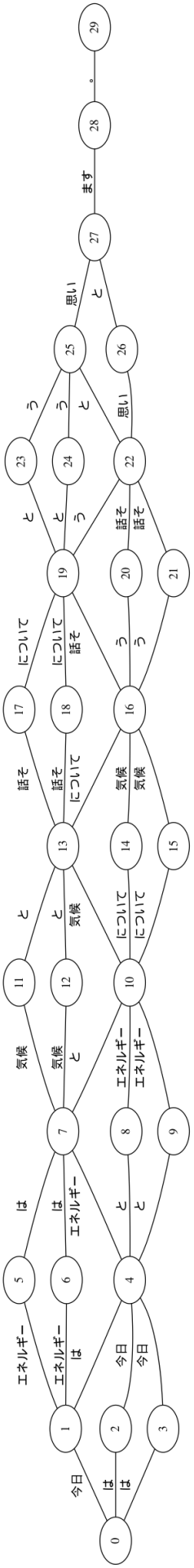

Figure 4.1: Reordering Lattice

5. Robot Implementation

Implementing the robot was done with help of the Institute of Professor Kobayashi. We got a simple Java framework to see how coding the robot is done. From there we started to think of different ways to implement detection of misunderstandings. Since the time was short the task was to get a simple demonstration of how this system could work. A simple working case to record a small demo (Table 5.2) video, which can be shown to people, was created.

5.1 Speech-recognition

One problem we faced was the lack of data for speech recognition of the lecture domain. The speech recognizer of Professor Kobayashis laboratory was optimized for running the Nandoku Word game, which is developed at their institute. So the recognizer was not able to recognize hesitations or questions asked by people.

We provided the laboratory with the data from Tatoeba corpus, so they could create a language model for their recognizer and recognize words and sentences out of the Tatoeba domain. The automatic speech recognition system called "SKOOD" is developed at the laboratory of Professor Kobayashi and is only able to perform well with around 20000 token language model. So we limited the Tatoeba corpus to the most occurring 20000 tokens.

After that we created a short video lecture about machine translation and showed it to students from the laboratory asking them to write down questions in Japanese. (Table 5.1) From those questions we build a second language model for recognizing words and sentences out of the machine translation lecture domain. Those language models were set for the recognizer with different weights so the recognition of the questions would become more probable than the sentences in Tatoeba.

We compiled a short demonstration as seen in Table 5.2. With this demo script we wanted to show some real life application of the robot translation system. In the left column we see the name of the person speaking, in the middle we have the actual spoken words and on the right we have the translation of the japanese for reading purposes. First the student interrupt the lecturer cause he wants to ask a question but hesitates, since English is not so easy for him. The robot detects the hesitations and offers translation services. The student then asks the robot to translate his question. After the answer of the lecturer the student hesitates again, cause he has a followup question. The robot again asks if translation is

needed. In the fourth part the student asks the robot directly to translate. After that the student has a simple question which he can ask by himself and the robot just observes. In the last part the student asks for translation again. The whole demonstration focuses on one slide of the small lecture we compiled and were conducted with some researchers of Professor Kobayashis laboratory.

English	Japanese
Could you say that again?	もう一度言ってもらえますか？
I am sorry I did not understand the triangle.	すみませんが、この三角形が理解できません。
How is this triangle called?	この三角形はなんと呼ばれているのですか。
Can you explain the triangle again?	この三角形をについてもう一度説明してもらえますか 。
Can you explain that again?	もう一度説明してもらえますか。
Can you explain slower please?	もう少しゆっくり説明してもらえますか。
Can you explain one more time?	もう一度説明してもらえますか。
Can you explain that in other words please?	他の言葉で説明してもらえますか 。
Whats happening inside the decoder?	デコーダーの中ではどのようなことをしていますか。
What do you do with unseen words?	未知の単語についてはどうしますか。
...	...

Table 5.1: Examples of the collected questions

5.2 Programming the Robot

Programming the robot and finding how to detect misunderstandings is a difficult task. In our case we thought of two different approaches.

5.2.1 Dectecting poor language skills by using language model score

A language model scores the probability that a given sentence is a valid sentence of a language. Hence the idea was to see if correct English sentences and spoken bad English sentences could be divided and recognized using the score of the TED or Tatoeba language model. Looking at different scores we found that the scores of the good sentences and the bad English sentences are too close together so it is not possible to find a good recognition of whether there is a problem or not.

5.2.2 Using hesitations and pauses to detect problems in conversation.

The second approach was to listen for pauses or hesitations in a conversation and if someone is hesitating too much or too long, ask if translation is needed. We wanted to achieve this to get the recognizer to be able to recognize Japanese hesitations. Since speech recognition is not perfect, some hesitations could be recognized wrong, so we implemented a second criterium for a hesitation might be that it is a string shorter than four syllables.

We also implemented some triggers that the robot can always be asked to translate. In hope to make the interaction with the robot feel more natural.

5.2.3 Integration

The integration of the translation system and the robot was the last part of our work. We had to connect to code of the robot and the machine translator. The translator was running as a web service on a machine at Karlsruhe Institute of Technology. Since the ports were still blocked at the time we used a ssh tunnel to forward the right ports to our web request from Waseda University. The web request would send a Japanese sentence to the translator and receive the translated English sentence.

After that the translated sentence would be passed to the speech synthesizer. In Figure 5.1 is a flowchart of the robot program which was implemented.

	Lecturer holds lecture	
Student	すみませんが、質問。。。	Excuse me, question...
Student	英語で何て言いう、えとー、あのー、何だっけ、。。。	How to say it in English, well..umm..
Robot	Shall I translate for you?	
Student	お願いします。	(yes) please
Student	トランスファとインタリングアのアプローチの違いは何ですか。	What's the difference between transfer and interlingua approach?
Pause		
Student	じゃ、お願いします	Please (translate I'm finished)
Robot	What's the difference between transfer and interlingua approach?	
	Lecturer answers	
Student	えとー、あのー、何だっけ。。。	Umm, well...
Robot	Shall I translate for you (again)?	
Student	お願いします。	(yes) please
Student	なぜインタリングアはこんなに複雑なのですか。	Why is interlingua so complicated?
Pause		
Student	じゃ、お願いします	Please (translate I'm finished)
Robot	Why is interlingua so complicated?	
	Lecturer answers again	
Student	シェーマもう一度お願いしてもいいですか	Shema, can you translate once again please?
Robot	I'm listening.	
Student	何かインラリングアのシステムはありますか？	Are there any interlingua systems?
Pause		
Student	じゃ、お願いします	Please (translate I'm finished)
Robot	Are there any interlingua systems?	
	Lecturer answers again mentions google	
Student	What system does google use?	
	Lecturer answers again	
Student	シェーマもう一度お願いしてもいいですか	Shema, can you translate once again please?
Robot	I'm listening.	
Student	ダイレクトトランスレシオンでは元の言語のルールは無理するということですか。	The direct approach ignores the rules of the language?
Pause		
Student	じゃ、お願いします	Please (translate I'm finished)
Robot	The direct approach ignores the rules of the language?	
	Lecturer answers again	

Table 5.2: Script for demonstration.

23

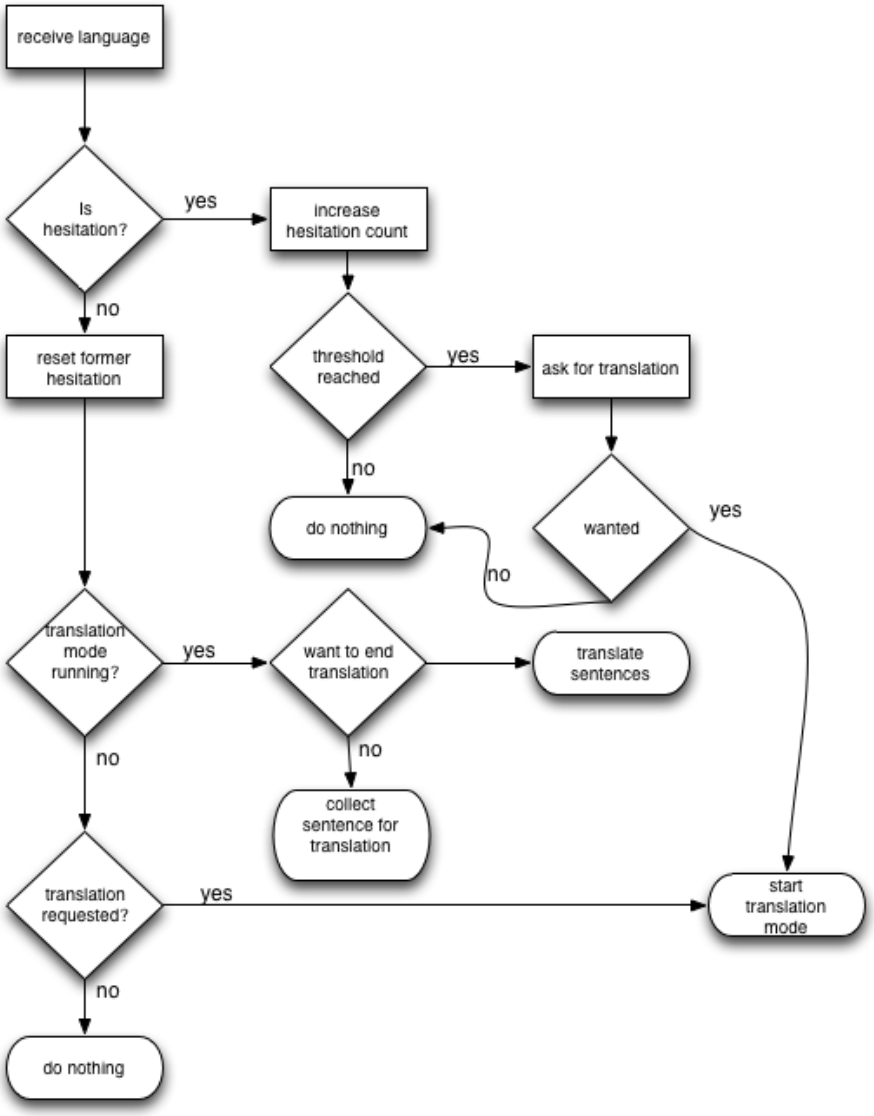

Figure 5.1: Conversation Robot Program Flowchart

6. Conclusion & Future Work

In this work we built a translation system for japanese to english translation using state of the art statistical machine translation tools. We concluded several experiments on the freely available TED, Kyoto Free Translation Task and Tatoeba corpora. Kyoto Free Translation Task data was only used to aid the alignments and probabilities since it has a very specific domain and we didn't consider it suitable for the task at hand. With these corpora we were able to achieve 16 BLEU points on Tatoeba corpus and around 4.6 BLEU on TED data. We concluded several experiments using a dictionary-based sequence-prediction method (MeCab) and a word-boundary prediction method (KyTea) for word segmentation, different phrase extraction heuristics, translation model adaption and reordering approaches. In the end we found to get the best translation scores using MeCab with JUMAN dictionary for tokenizing, the grow-diag-final-and heuristic for phrase extraction using only in-domain data (Tatoeba) and Long-Range POS based reordering approach. We also developed a simple system for a humanoid robot moderator, which is able to recognize simple hesitations and interact with the participants. This robot software was included in the framework of Professor Kobayashis laboratory. Together with this framework we created a demonstration of the humanoid robot moderation system for lecture translation context.

Since we just built a basic version of the whole framework we want to point out the main problems that still exist and need to be solved in future work. One question that arises is, why build a robot that aids the translation and not just use a computer that will translate when triggered?

In further research when the robot would actually be able to detect flaws and misunderstandings and be able to follow conversations and understand the topic that is talked about. It would be a device which is able to improve conversation significantly since it could translate in several different languages and make sure that everybody got a right understanding of the topic. Interaction with the robot would feel natural and not disturb the flow of conversation. Right now the system needs more work and there are a lot of problems that are yet still unsolved.

The speech recognizer was limited to around 20000 words. Newer systems can handle much more than that and thus probably improve the recognition performance. Which is one of the most important tasks in this system. If the system is not able to understand what the speaker was saying there is little chance that the translation will be good. Right now speaker independent speech recognition of continuous speech is still and ongoing research.

Also different forms of background noise are problematic and decrease speech recognition performance drastically.

Since the system only recognized misunderstandings by counting the hesitations it is very prone to errors. More sophisticated systems could try to track topics. Once the interacting people change the topic while speaking in the foreign language it could be a clue to a misunderstanding or communication problem. This is also field of ongoing research and a very hard task todo, that is why we decided to limit my approach to simple hesitation counting. The robot was only able to recognize a few prepared sentences which would trigger or stop the robot interaction. So you would have to tell the robot directly that it is spoken to. The number of sentences could be increased or some heuristics implemented which try to get the meanings of generic sentences so the robot interacts more natural in conversation. Also the movement of the robot was very limited. It was always looking at the current speaker and nodding its head 60 percent of the time to give the speaker a sense of understanding, which wasn't perceived very natural, though no user studies have been conducted on that. The translation from Japanese to English needs further improvement to generate good translations on a broader domain. The system is able to translate sentences within the domain very well but has problems at generalizing to generic sentences. We guess TED has more potential than we were able to discover given the time constraint. Since the reordering rules did not look very promising some parameters on that model have to be tweaked to get a better result. Also we want to try to use some discriminative word alignment to see if better alignment will improve translation qua lity of the TED data.

Bibliography

[KBC08] J. Katz-Brown and M. Collins, "Syntactic reordering in preprocessing for japanese to english translation: Mit system description for ntcir-7 patent translation task," 2008.

[KHB+07] P. Koehn, H. Hoang, A. Birch, C. Callison-burch, R. Zens, A. Constantin, M. Federico, N. Bertoldi, C. Dyer, B. Cowan, W. Shen, C. Moran, and O. Bojar, "Moses: Open source toolkit for statistical machine translation," in *Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, June 2007, pp. 177–180.

[KOM03] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, ser. NAACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 48–54. [Online]. Available: http://dx.doi.org/10.3115/1073445.1073462

[LMP01] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289. [Online]. Available: http://dl.acm.org/citation.cfm?id=645530.655813

[LO05] C. Lioma and I. Ounis, "Deploying part-of-speech patterns to enhance statistical phrase-based machine translation resources," in *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ser. ParaText '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 163–166. [Online]. Available: http://dl.acm.org/citation.cfm?id=1654449.1654486

[Neu11] G. Neubig, "The Kyoto free translation task," http://www.phontron.com/kftt, 2011.

[NK+09] J. Niehues, M. Kolss *et al.*, "A pos-based model for long-range reorderings in smt," *Fourth Workshop on Statistical Machine Translation (WMT 2009), Athens, Greece*, 2009.

[NNM11] G. Neubig, Y. Nakata, and S. Mori, "Pointwise prediction for robust, adaptable japanese morphological analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 529–533. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002736.2002841

[ON00] F. J. Och and H. Ney, "Improved statistical alignment models," in *ACL00*, Hongkong, China, October 2000, pp. 440–447.

[RV07] K. Rottmann and S. Vogel, "Word reordering in statistical machine translation with a pos-based distortion model," 2007.

[Sto02] A. Stolcke, "Srilm - an extensible language modeling toolkit," 2002, pp. 901–904.

[VZH+03] S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venogupal, B. Zhao, and A. Waibel, "The CMU statistical translation system," in *Proceedings of MT Summit IX*, New Orleans, LA, September 2003.